Improving e-book access via a library-developed full-text search tool* EC

Jill E. Foust, MLS; Phillip Bergen, MA, MS; Gretchen L. Maxeiner, MA, MS; Peter N. Pawlowski, BS, BA

See end of article for authors' affiliations.

Purpose: This paper reports on the development of a tool for searching the contents of licensed full-text electronic book (e-book) collections.

Setting: The Health Sciences Library System (HSLS) provides services to the University of Pittsburgh's medical programs and large academic health system.

Brief Description: The HSLS has developed an innovative tool for federated searching of its e-book collections. Built using the XML-based Vivísimo development environment, the tool enables a user to perform a full-text search of over 2,500 titles from the library's seven most highly used e-book collections. From a single ''Google-style'' query, results are returned as an integrated set of links pointing directly to relevant sections of the full text. Results are also grouped into categories that enable more precise retrieval without reformulation of the search.

Results/Evaluation: A heuristic evaluation demonstrated the usability of the tool and a web server log analysis indicated an acceptable level of usage. Based on its success, there are plans to increase the number of online book collections searched.

Conclusion: This library's first foray into federated searching has produced an effective tool for searching across large collections of full-text e-books and has provided a good foundation for the development of other library-based federated searching products.

Highlights

- Electronic Book Search searches the content of over 2,500 e-books contained in the 7 most popular packages in the HSLS collection.
- A good federated search tool is founded on a thorough understanding of the target products, careful configuration, and regular monitoring.
- Clustering technology is a useful means of narrowing the large results sets often associated with keyword searching.

Implications

- Users need more efficient ways to access the content of a library's e-book collection.
- Federated search engines targeting specific collections or user populations in a library can be effective tools.

INTRODUCTION

The emergence of the Internet has introduced new ways for users to access library resources and has shaped user behavior and expectations [1]. Users now expect instant and constant access to information, often from distant locations, and as a result, remote access to online library resources has become an increasingly significant part of library service.

Remote access has become especially important at the University of Pittsburgh's Health Sciences Library System (HSLS). In addition to supporting the university's schools of the health sciences, HSLS also supports the 17 hospitals of the University of Pittsburgh Medical Center (UPMC). Medical staff members at many of these outlying facilities lack access to a physical library, so online resources, particularly online reference works, are of especial value. HSLS offers remote access to a vast collection of electronic materials, including over 3,000 licensed electronic books (e-books) that are represented in the library's online catalog and also in a Web-based alphabetical title list. These materials should be especially useful for those seeking quick information in the course of patient care; however, the necessity of identifying relevant online titles and then searching within each title separately slows the information retrieval process and potentially limits e-book use [2].

To address this problem, HSLS chose to create a federated search tool called Electronic Book Search <http://www.hsls.pitt.edu/booksearch>. Built using Vivísimo's Velocity software package, this tool facilitates use of the library's e-book collection by enabling users to search the full text of a large set of e-books from across the collection in one easy step. This paper describes the process of developing and implementing

* Based on a presentation at MLA '05, the 105th annual meeting of the Medical Library Association; San Antonio, Texas; May 16, 2005.

EC Supplemental figures are available with the online version of this journal.

Electronic Book Search and the challenges and successes encountered along the way.

BACKGROUND

The federated search tool has emerged as a successful means of meeting the information needs of users despite several limitations. From a streamlined interface with a single search, users can pull results from a variety of sources, both public and subscription-based [3–7]. There is the additional benefit that some of these results may be from resources the user might not have otherwise found [3]. The speed and ease of federated search tools make them appealing to searchers, particularly novice searchers. They do not need to learn how to formulate searches in each of the databases, nor must they learn how to interpret the different results sets; federated search tools typically display results in a common format [4]. An efficient federated search engine allows searching to be done in a timely manner and eliminates the need for a user to learn multiple search interfaces [5].

There are trade-offs for this simplicity, however. The search capabilities of the federated search are limited by those of the individual databases or sources. Boolean searching is not usually possible because a federated search engine can only process what the target provider allows [6]. Advanced search commands such as field searching and truncation are also not usually possible. As a result, information literacy may suffer due to the abandonment of traditional search skills in favor of a ''Google-style, keyword approach'' [5].

Two issues also arise in the integrated results. First, federated searches commonly retrieve duplicate results from different sources. Because sources return results in small sets (of typically ten to twenty), the federated search tool's deduplication process has only a small portion of the results to work with at a time [7]. A complete deduplication would require that all search results be compared. With a large retrieval, this would be extremely time-consuming, thus not feasible for most searchers. Second, federated search engines do not perform relevancy ranking well. While content providers can utilize the full article and its indexing, federated search engines have only the citation with which to work [7]. Thus, federated search engines have only limited data available on which to base relevancy, and this data may be insufficient.

Although federated searching may seem simple to the user, the set-up of a tool can be complicated. Individual sources store and present data differently, requiring behind-the-scenes efforts to make and keep results compatible. The mapping of data between the sources and the search tool can be quite complex and challenging to maintain, because sources can change their data format and output at any time [5]. Subscription-based sources also present difficulties, since licenses with the library define permissible user groups [4]. Authentication processes need to be established to allow valid users access to the resources while unauthorized users are excluded.

While the literature does not contain a discussion of federated searching used with full-text e-books specifically, the more general literature confirmed the library's decision to apply federated search technology to this situation and offered insight as to where future problems might occur.

ELECTRONIC BOOK SEARCH DEVELOPMENT

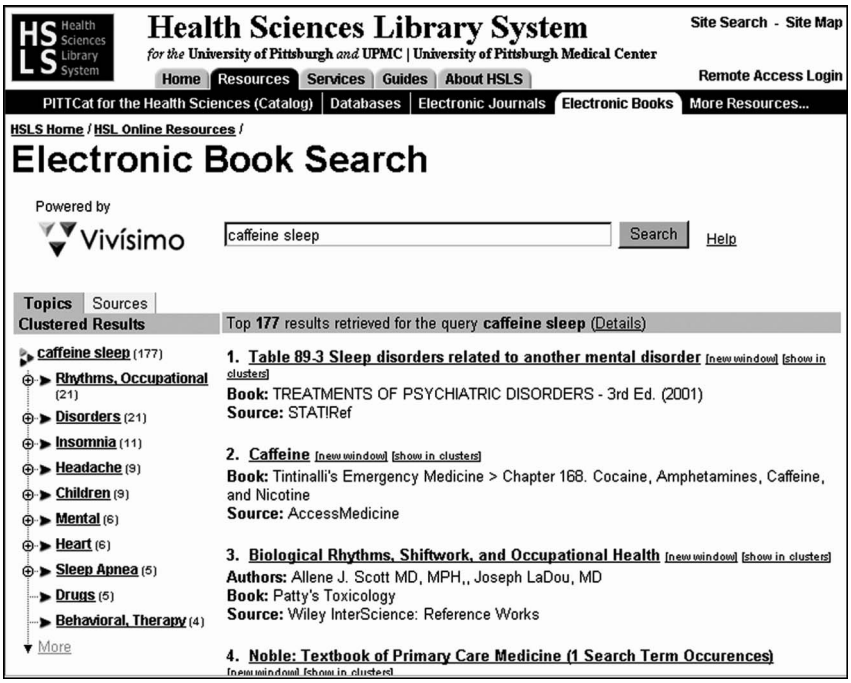
Setup

Though there are a number of federated search products on the market, HSLS did not explore these products because the organization had recently licensed Vivísimo's Velocity, an XML-based development environment that consists of a set of software tools for building information retrieval applications. It was decided that this project presented a good opportunity for the first locally developed application using the software. Velocity comprises three interrelated but distinct tools: (1) the Enterprise Search Engine, which allows for automated indexing and searching of document collections (not utilized in this project); (2) the Content Integrator, which transmits queries to the Web-based search engines of individual data sources and integrates the results from each into a single result set; and (3) the Clustering Engine, which takes a result set and dynamically groups it into meaningful categories, placing results in hierarchical folders. Clustering can be a useful way of managing large results sets, and may be familiar to readers from Clusty, Vivísimo's Web-based metasearch engine, or ClusterMed, its subscription-based product that searches and clusters PubMed [8, 9].

The development team, consisting of the Information Architecture Librarian, the Web Manager, who is also a Reference Librarian, and the Cataloging Librarian, selected packages of licensed e-books from 7 different vendors for inclusion: AccessMedicine, Books@Ovid, ebrary, Elsevier ScienceDirect, MD Consult, STAT!Ref, and Wiley InterScience. These were selected for the quality and number of titles in each, and also given the expectation that HSLS would continue to license them over time. Overall, these packages present over 2,500 e-books, a vast majority of the HSLS e-book collection, and include all of the most popular e-book titles.

A detailed profile identifying functionalities and specifications was prepared for each package, or target product, to allow for the proper configuration of the Vivísimo Content Integrator. All of the products include a Web-based search engine for accessing the content of their titles, which is required for interaction with the Content Integrator; however, each of the packages features a unique set of characteristics requiring special configuration. Some of the characteristics that had to be examined include: (1) Searching: What are the basic and advanced search features? Does it search the full collection or only subscribed titles? Does it search full text, book indexes, or other data? (2) Results: Does it sort results by relevancy? What information does it provide about each result? How many

Figure 1
Electronic book search results screen



results does it return and in what grouping? (3) Authentication: How does the e-book provider recognize valid users? How does it maintain a user's session? (4) Data transmission: Since the Content Integrator does not use the Z39.50 interface, what is the HTML format of the results data? What are the parameters of its common gateway interface (CGI) program?

Utilizing these identified characteristics, the initial stage of programming focused on configuring the Content Integrator to interact with each of the targeted products. For each provider, a ''source'' file was created in Vivísimo to store the information necessary for the Content Integrator to translate queries into the appropriate format for the provider's CGI program, send the query to the CGI program, and interpret the returned results. This information includes the CGI program's URL and parameters, the syntax the search engine supports, and the format of the HTML that is returned. The setup for each source was tested for errors and unexpected results, and after fine-tuning, the sources were linked together for federated searching.

Display

Attention then turned to the integrated results display. Ideally, results, regardless of their source, should be presented in a consistent format throughout the final set, but the reality is that each target product returns different information about results and uses different formats. In order to make providers' results more compatible, the team reassessed which data elements from each would be displayed and defined the field labels. The maximum number of results was also determined by trial and error, set to be large enough to allow for effective clustering but modest enough to not unduly slow retrieval time. An overall maximum set of 450 results was chosen, divided amongst the providers. The results are displayed based on general relevance: each provider offers its results ranked by relevancy, and these are interfiled in the default Vivísimo display.

The basic results screen contains this list of resulting citations and also the clusters provided by the Clustering Engine (Figure 1). Without any special configuration, this component of the Velocity package uses the information in the results set to establish dynamic categories and sorts the results into hierarchical folders. Although customization is possible, the default settings were considered to be successful in initial testing and so customized capabilities were not explored. Only a short set of clustering stop words was defined. For example, words such as ''Chapter'' and ''Introduction'' are common in the results but do not provide useful categories for clustering; thus, these were excluded from the clusters. Clustering by topic and by ''source'' are included; this latter feature presents results on a provider-by-provider basis.

The development team was initially unsure how much explanation and guidance would be required by users of the Electronic Book Search. Clustering in particular would likely be a new concept to most. However, following Vivísimo's lead in their ClusterMed product, a minimalist approach was adopted. Instruction is provided in the form of a diagram on the opening screen (Figure 2; online only) that explains the different components of the results screen. A simple Help section also offers general information about the tool, tips on search syntax, a list of the included e-book titles, and an email link for requesting assistance.

Testing

Electronic Book Search was subjected to two types of testing prior to its release. First, all HSLS librarians were asked to test it. Only a minimal amount of information about the product was given so as to simulate the typical user situation. These testers were asked to evaluate the product from the perspective of both the librarian and the patron, and the resulting feedback was positive. Testers found the tool to be a timesaver because they could search numerous titles simultaneously and because they could explore multiple e-books in situations in which they would not know which titles to search. The clustering was useful in pulling out different aspects of the search topic, for example, pediatric coverage of a particular disorder. Finally, most felt that there was appropriate guidance and help available for patrons to use the product successfully.

Second, HSLS employed a heuristic evaluation in which evaluators compared aspects of the site against a set of known usability principles [10, 11]. Heuristic evaluation does not test how well a product will meet users' needs; rather, it identifies design problems that would adversely impact a user's interaction with the product. Five HSLS librarians and one external information professional served as evaluators. Working individually to assess the product, they identified a total of twenty-five different problems and then ranked them for their level of severity. Only one of the twenty-five was rated by the evaluators as a major usability problem, that is, ''important to fix'': inadequate feedback that a search is in progress. Because searches were slow and nothing was happening on the screen, users might not have realized the system was running. This problem was immediately addressed by adding brightly colored text above the search box indicating that a search is in progress. The remaining usability problems identified in the evaluation, all ranked as minor or cosmetic, were then dealt with in order of severity and feasibility. HSLS has not yet undertaken any testing with the product's primary users, the library's patrons. Such testing, however, will likely become part of the evaluation process for future iterations now that basic usability has been addressed.

DISCUSSION

Success

Overall, the development of Electronic Book Search has been a success. Informal user feedback offered during reference service interactions and through departmental liaisons has been favorable, frequently indicating that the tool allowed users to quickly find the answer or resource they were seeking.

Usage of the Electronic Book Search was charted, employing WebTrends software to track visits over an 8-month period, from March 2005 through October 2005 (Figure 3; online only). In that time, a total of 2,008 visits were made to the system. After the initial release to the public in early 2005, there was low use of the product in March with 145 visits. This increased dramatically in April to 407 visits after it was publicized on the HSLS home page and in the library's print and online newsletter [12]. Usage decreased after this peak and has been inconsistent in the months following. However, with an average of 251 monthly visits during the test period, HSLS is satisfied with the overall level of use, which is expected to increase following a current Website redesign project, which will display the tool more prominently on the site.

In addition to numerical data, WebTrends also tracks the exact search terms used in queries. Electronic Book Search is designed for ''Google-style'' queries, that is, simple words or phrases. During the development phase, there was a concern that this type of searching would not suffice for users in the medical community, who may expect advanced search capabilities. The web server log analysis suggests that this is not the case. Of 159 unique search terms entered in March 2005, shortly after the public release of the tool, 90% present the expected keyword or phrase format. Only 10% of the searches contain complex Boolean strings, structured search phrases, or attempts to search for book titles or authors rather than content. This would suggest that most users did not expect advanced search capabilities and easily grasped the type of search that Electronic Book Search expected.

Although the project focused on the federated searching capabilities of the Vivísimo software, its clustering capabilities proved to be an added benefit. Keyword searching can yield large numbers of results, some of which are likely to be irrelevant to the user. Instead of reformulating additional searches that are more restrictive, the user can take advantage of Electronic Book Search's clusters to focus on the most appropriate hits. Thus, as seen in Figure 1, a user searching on *caffeine sleep* who is daunted by the 177 results can look to the clusters in order to hone in on results with a pediatric context or those discussing headaches. In both instances the tabs can be expanded to further narrow the results set. This capability has rated well in the informal feedback HSLS received about Electronic Book Search and has contributed to its success.

Challenges

Several of the ongoing challenges for federated search engines that are identified in the literature are likewise present for this project. For example, the system does not readily accept advanced search commands. This includes field-specific searching, truncation, and complicated Boolean search strings. However, because all of the selected sources accommodate at least simple Boolean search, the user does have this option in Electronic Book Search.

Programming can also be complicated for those e-book providers that do not automatically map a search to the full set of subscribed titles. A source may require the manual selection of each title to be searched; at the other end of the spectrum, if the source automatically searches its full suite of resources (even if the user will not have access to all results in full text), one may have limit the search to particular titles. In

both cases, programming is done on a title-by-title basis and requires monitoring in case the subscription contents change. At HSLS, the cataloger responsible for e-book cataloging reports these changes as part of her regular workflow.

The system will also require regular monitoring to ensure that no changes to the target products' CGI programs have occurred that would block communication between Electronic Book Search and the sources. Although the system only includes seven products, two were redesigned in the first six months that Electronic Book Search was available. Since both revisions were advertised, preparations could be made for reprogramming the corresponding Content Integrator source file, although in fact little could be done before the interface redesigns went live. Not all changes are advertised, however. One provider, for example, added a temporary welcome screen advertising upcoming enhancements; while this seems innocuous enough, it effectively blocked access to this provider's resources via Electronic Book Search until the Electronic Book Search setup was reprogrammed.

Working with licensed resources presents a variety of challenges. For example, it is possible to exceed the maximum number of concurrent users as allowed by the provider license. In the case of Electronic Book Search, each search opens a session on each source's search engine, even if the user does not choose to view that source's full-text results. This has caused an increase in the number of open sessions for the included providers. So far, this has not been an issue, but it is a situation that must be monitored as use of Electronic Book Search increases. There are also situations in which special arrangements with the provider must be made. In one case, a provider added an individual license agreement screen to which a user must respond before accessing the provider's resources; because this interfered with Electronic Book Search, HSLS contacted the provider for permission to bypass the page, which was granted.

An issue that had not been anticipated from the literature review is that of slow search speed due to the large number of e-books that are searched concurrently. While a federated search is certainly faster than searching each of the component sources separately, Electronic Book Search runs more slowly than some users might expect. This is addressed in part with on-screen feedback to indicate that a search is in progress. However, latency remains a concern and may be a factor whenever the addition of further e-book collections is considered, since those additional collections will further slow search time.

Fortunately, two of the challenges frequently cited in the literature proved to be non-issues for Electronic Book Search. First, the duplication that tends to be a problem in federated searching occurs when the target products overlap in their search coverage and potentially retrieve the same results; for example, if they are searching the web and return the same URL. Deduplication is not a worry in this case since each provider is searching only its own set of resources. Potentially there could be redundancy if the same work is included in multiple e-book packages, since federated search engines identify duplicates based on URL, not on the content. However, this is not a concern for this project because HSLS has virtually no overlap of titles in the different providers' collections. Secondly, relevancy among the integrated search results does not present a problem. All of the sources used in Electronic Book Search rank their own results sets, and Vivísimo's Content Integrator simply interfiles these so that a relative relevancy is achieved. This approach seems to be satisfactory.

CONCLUSION

This paper has described one medical library's foray into federated searching, in this case for the development of a tool that allows users to search a large collection of e-books simultaneously across providers and at a full-text level. The simple interface and ability to search across multiple resources at one time make Electronic Book Search appealing to users. Additionally, feedback and evaluations demonstrated a sufficient level of satisfaction and use to continue maintaining this tool despite its inherent ongoing challenges.

Electronic Book Search could potentially impact users from outside the HSLS community as well. Although Electronic Book Search was not deliberately designed for external use, the setup allows anyone with Internet access to conduct searches on the contents of the HSLS e-book collection and to view the results lists. Licensing restrictions prevent users without recognized subscriptions for the individual resources from viewing the full text of the results. However, if users outside the HSLS community have subscriptions to the resources either individually or through their home institutions, they will be able to follow the links in the results list to the full text of those e-books. Although other medical libraries may not subscribe to all of the same titles, it is likely that their e-book collections will closely mirror that of HSLS, thus allowing their patrons to benefit as well from the improved access to these electronic resources.

In fact, in response to the successes of Electronic Book Search, HSLS has since explored other library applications for Vivísimo's Velocity software. One such application, available through the HSLS Molecular Biology and Genetics Web site <http://www.hsls.pitt.edu/guides/genetics>, provides access to over 900 online bioinformatics databases and software tools. With this, users can locate appropriate resources more efficiently than with the standard popular Web search engines. Other future applications are also being considered as HSLS continues to find innovative ways of addressing the information needs of its users in the online environment.

REFERENCES

1. Covey DT. The need to improve remote access to online library resources: filling the gap between commercial vendor

and academic user practice. Portal Libr Acad 2003;3(4):577–99.
2. Coiera E, Walther M, Nguyen K, Lovell NH. Architecture for knowledge-based and federated search of online clinical evidence. J Med Internet Res [serial online]. 2005;7(5):e52. [cited 21 Jun 2006]. <http://www.jmir.org/2005/5/e52/>.
3. Stewart VD. Federated search engines. MLA News 2006 Jan:17.
4. Fryer D. Federated search engines. Online 2004 Mar/Apr; 28(2):16–9.
5. Curtis AM, Dorner DG. Why federated search? Knowl Quest 2005 Jan/Feb;33(3):35–7.
6. Wadham RL. Federated searching. Libr Mosaics 2004 Jan/Feb;15(1):20.
7. Hane PJ. The truth about federated searching. Inf Today 2003 Oct;20(9):24.
8. Markoff J. New company starts up a challenge to Google. NY Times 2004 Sep 30; Sect. C:6 (col. 6).
9. Price G. Reducing information overkill. SearchDay. [Web document]. 2004 Sep 30. [cited 29 Mar 2006]. <http://searchenginewatch.com/searchday/article.php/3415071>.
10. Nielsen J. How to conduct a heuristic evaluation. [Web document]. [cited 29 Mar 2006]. <http://www.useit.com/papers/heuristic/heuristic_evaluation.html>.
11. Nielsen J. Ten usability heuristics. [Web document]. [cited 19 Sep 2006]. <http://www.useit.com/papers/heuristic/heuristic_list.html>.
12. Bergen P, Maxeiner G. Electronic Book Search simplifies full-text searching. HSLS Update 2005 Feb;10(1):1,3. [cited 21 Jun 2006]. <http://www.hsls.pitt.edu/about/news/hslsupdate/2005/february/vivisimo_ebook_search>.

AUTHORS' AFFILIATIONS

Jill E. Foust, MLS, jef2@pitt.edu, Web Manager/Reference Librarian; **Phillip Bergen, MA, MS,** bergen@pitt.edu, Information Architecture Librarian; **Gretchen L. Maxeiner, MA, MS,** maxeiner@pitt.edu, Cataloging Librarian, Health Sciences Library System, Falk Library of the Health Sciences, University of Pittsburgh, Pittsburgh, PA 15261; **Peter N. Pawlowski, BS, BA,** pawlowski@vivisimo.com, Software Engineer and Lead Linguist, Vivísimo, Inc., 1710 Murray Avenue Suite 300, Pittsburgh, PA 15217

Received March 2006; accepted September 2006